Data Integration and Predictive Analysis System for Disease Prophylaxis

Madhav Erraguntla
KBSI
merraguntla@kbsi.com

John Freeze
KBSI
jfreeze@kbsi.com

Dursun Delen
OSU
dursun.delen@okstate.edu

Karthic Madanagopal
KBSI
kmadanagopal@kbsi.com

Ric Mayer
KBSI
rmayer@kbsi.com

Jam Khojasteh
OSU
jam.khojasteh@okstate.edu

Abstract

The goal of the Data Integration and Predictive Analysis System (IPAS) is to enable prediction, analysis, and response management for incidents of infectious diseases. IPAS collects and integrates comprehensive datasets of previous disease incidents and potential influencing factors to facilitate multivariate, predictive analytics of disease patterns, intensity, and timing. IPAS supports comprehensive epidemiological analysis - exploratory spatial and temporal correlation, hypothesis testing, prediction, and intervention analysis. Innovative machine learning and predictive analytical techniques like support vector machines (SVM), decision tree-based random forests, and boosting are used to predict the disease epidemic curves. Predictions are then displayed to stakeholders in a disease situation awareness interface, alongside disease incidents, syndromic and zoonotic details extracted from news sources and medical publications. Data on Influenza Like Illness (ILI) provided by CDC was used to validate the capability of IPAS system, with plans to expand to other illnesses in the future. This paper presents the ILI prediction modeling results as well as IPAS system features.

1. Introduction

Diseases are a constantly changing threat to public health. Each year, certain diseases such as influenza resurface in seasonal outbreaks. Other diseases make a startling rise in infected individuals in certain unexpected locations, as is the case for Zika or Ebola in the last few years. In either situation, spread of these diseases may be preventable with appropriate disease-management measures. Advance knowledge about the location, timing, peak intensity, and potential number of infected will help public health stakeholders in taking proactive disease containment and management efforts [1, 2, 3]. Health organizations such as the CDC have recognized this fact, and have sponsored several competitions and workshops to encourage development of viable prediction models [7, 8, 13].

Effective disease-management activities need accurate prediction of disease outbreaks. The timely prediction of a disease outbreak facilitates the effective coordination and mobilization of medical, human, and pharmaceutical resources. Prior knowledge of potential disease occurrences enables proactive development of medical interventions, medical prophylaxis to disease hazards, and containment of disease vectors. Traditional epidemiology has focused on compartmental models (susceptibility, exposed, infected, recovered (SEIR) based approaches for forecasting disease progression [4, 5, 6]), which are data and modelling intensive. IPAS takes advantage of innovative machine learning and predictive analytics to facilitate creation of generic disease prediction models.

Others seek to analyze previous disease emergence to predict future emergence. Investigations have attempted to link disease outbreaks to weather [14], bushmeat consumption [9], socioeconomic status [10, 11], and a breadth of other factors. Disease occurrence predictions like these rely upon the collection and curation of quantitative as well as qualitative data related to historical disease occurrences, weather and environmental data, and vector data (see Figure 1). Recognizing the importance of data collection, researchers like HealthMap [12] and VectorMap [13] are collecting and organizing data related to diseases, vectors, and environmental conditions; however, these datasets are currently isolated. IPAS collects and integrates the relevant disease data to facilitate comprehensive epidemiological analysis. Natural language processing is used to extract specific disease, syndromic, and zoonotic details from news feeds, public health reports, medical publications and social media. The extorted data is used in disease analysis and creation of biosurveillance reports (Figure 2).

URI: http://hdl.handle.net/10125/41287
ISBN: 978-0-9981331-0-2

Collection of data from different sources will lead to issues with data quality, inconsistent and incompatible data, and mismatched terminology [23]. Producers of these data perform ad-hoc, proprietary, behind the scenes data manipulations that are neither transparent nor scalable. IPAS provides fundamental approaches to data cleaning, and spatial and temporal harmonization to allow researchers to integrate and analyze the collected data. Supported core processing steps include spatio-temporal clustering, correlation analysis, and determination of the factors influencing the spread of a disease.

IPAS leverages machine learning and predictive analytics based techniques such as kernel-based support vector machines, nearest neighborhood calculation, decision trees, random forests, and boosted trees for the purposes of disease incidence prediction [9, 10]. This paper presents the results of application of these analytical models to predict the ILI within the USA and Health and Human Services (HHS) regions.

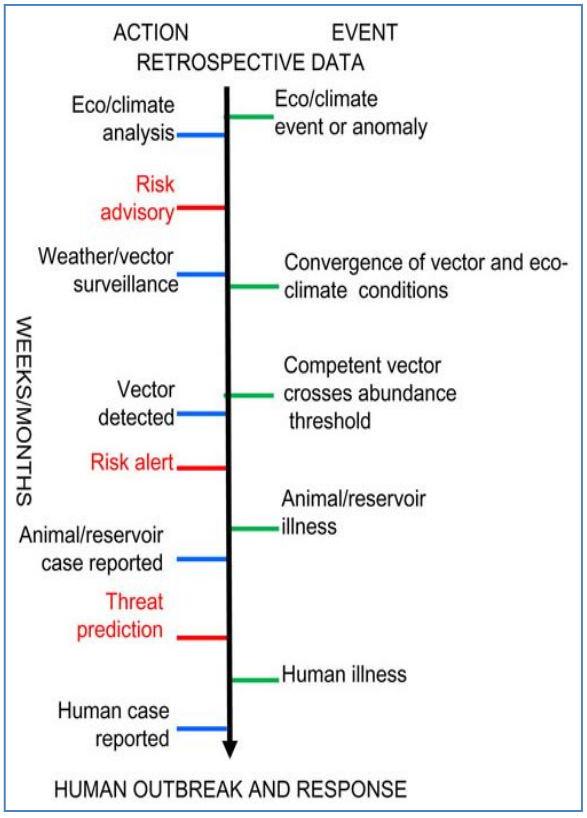
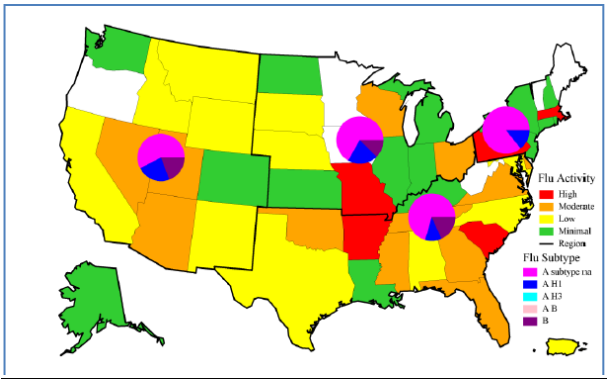


Figure 1. Need For Integrating Multiple Indicators for Effective Predictive Models[1]



In NORTHCOM during week 02[1,2]

- Influenza activity was low to moderate for most states.
- The percentage of outpatient visits due to ILI continued to decreased from a peak at week 52, but remained above baseline.
- Flu A and B viruses continued to circulate with similar counts of A/H1 and A/H3 viruses identified.
- The percentage of positive lab tests for week 02 was 5.5% for service members and 5.4% for beneficiaries.
- Two influenza hospitalizations (RME) among vaccinated service members were reported for week 01.

Figure 2. Sample Influenza Biosurveillance Report

After models generate predictions for disease outbreaks, the information must be effectively disseminated to stakeholders, decision makers, and public health officials. Currently, biosurveillance analysts at the CDC, DoD and WHO create and share weekly disease biosurveillance reports to summarize and communicate the state of the focus disease. These reports summarize the disease, reported incidents, and detect patterns and observations (see Figure 2). While this activity involves analysis and insights from the epidemiological experts, most of the analyst time is spent on collecting and aggregating the necessary information. IPAS automates the data collection and aggregation, thereby reducing the burden on biosurveillance and epidemiology analysts currently spending thousands of hours each month on this effort.

2. IPAS Solution Dataset

Influenza Like Illness (ILI) was selected as the focus syndrome for predictive analytics for the IPAS solution. Historical ILI data was provided by CDC through the FluView website [16]. The data consisted of the percentage of patients visiting medical care facilities with ILI in the United States. Data are available from 1997, but prior to 2002, data for off-season weeks is missing. Data are provided for 10 HHS regions, defined

by the US Department of Health and Human Sciences (USDHSS) and at the national level. Percentages of lab specimens that tested positive for virus types A and B, as well as vaccination rates in various HHS regions are also provided. The goal of the flu prediction effort was to predict the future flu trajectory based on current and historical data. Specifically, the Week 1, 2, 3, and 4 look ahead predictions, season start time prediction (season start is defined when the ILI values are above threshold for three consecutive weeks), and peak ILI value and timing.

As part of data collection and integration, we incorporated environmental data from the National Oceanic and Atmospheric Administration's (NOAA) Global Historical Climatology Network (GHCN). This data set includes recorded data from weather stations across the United States. Data about minimum and maximum temperature, snowfall, snow depth, precipitation, and pan evaporation (which is a measure of relative humidity and temperature) on a daily level is available in this dataset. As part of data cleaning and transformation, we aggregated the data for HHS regions and the national level, and included average minimum and maximum temperatures, average precipitation, average snowfall, and average pan evaporation as potential influencing environmental variables in the prediction model.

The social media data was drawn from the HealthTweets research performed by Johns Hopkins Social Media and Research Group. These data are available for the flu years 2012-2014, and give the raw and normalized amount of Tweets related to influenza per week by state. The signal generated from the Tweet data related to ILI was incorporated into an additional prediction model. Due to limited data, model training and testing processes utilized the same data. We acknowledge the folly of this technique, and as such the data for these predictions is not presented here. The resulting predictions with the social media prediction model showed worse performance than the original predictions.

In addition to the directly available ILI data, a number of derived variables were created for inclusion in the prediction model using MATLAB. These variables were created to extract the pattern of ILI, introduce non-linearity, and account for the time series nature of the data. The derived variables include:

- the percentage ILI change over the last two weeks
- percentage ILI change over the last three weeks
- cumulative ILI for the flu year
- cumulative ILI for the last four weeks
- cumulative ILI for the last eight weeks
- ratio of specimens testing positive for strains A & B
- square and cube of ILI
- square and cube of cumulative ILI for the flu year
- inverse of ILI (for start/peak timing predictions)
- number of consecutive ILI increases
- number of ILI increases above a threshold value

The derived variables were calculated and written to a comma-separated file (.csv). For data involving dates, the date was converted to a week number according to HHS standards. Each week is a seven-day period starting on Sunday and ending on Saturday. If January 1st is on a Thursday, Friday, or Saturday, then all days of that week remain a part of the previous week. According to these rules, January 4th will always fall on the first week of the year. Due to these standards, each year contained either 52 or 53 weeks. Additionally, each flu-year is considered to start on the 40th week and end on the 39th week of the calendar year; for the purpose of variable definition, the 40th calendar week was renamed to flu-week 1, constraining the flu-week variable to increase monotonically for any given flu-year.

3. Prediction Modeling

The rich set of integrated ILI and environmental data were used to develop the prediction model. We developed non-linear regression, decision tree-based boosting, and support vector machine (SVM) based machine learning models for prediction using packages available within R. Models were trained using 10 years of data (2002-2011) and validated on three years (2012-2014). The model performance was evaluated only on the validation data. Variables that had less than significant influence on prediction were excluded from the final model. All of the predictive models performed similarly, with SVM having a slightly better performance for 1-week look ahead predictions (see Table 1). Time limitations restricted our investigation to only one of the three available methods; SVM was chosen as the predictive model to focus the model development effort although the other options are expected to produce similar results.

Table 1. Comparison of Performance of Different Prediction Models for 1-Week Look Ahead

Prediction Model	Average Mean Square Error on Validation Data
Regression	0.1439
Boosting	0.1428
SVM	0.1401

1-, 2-, 3-, and 4-week look-ahead predictions from the SVM model for each week are presented in Figure 3. As can be expected, one-week predictions are more accurate than longer lead-time predictions, with 4-week predictions having more error rates. The prediction

model was able to efficiently capture the cyclical nature of the ILI with a R^2 of 90%.

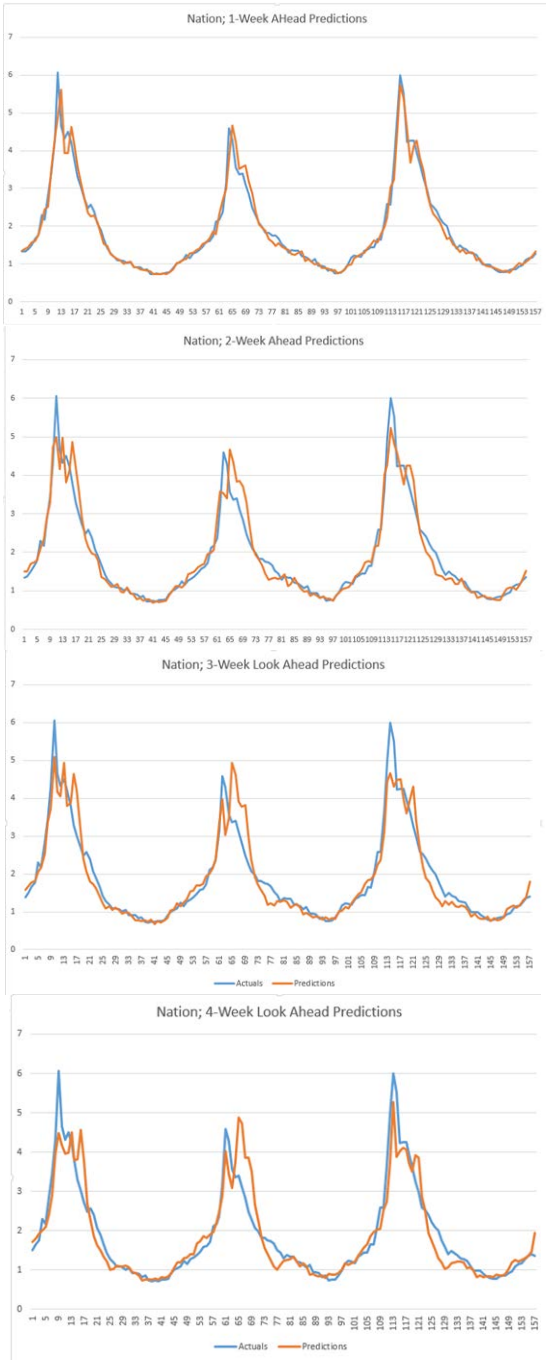


Figure 3. National Level ILI Predictions for 2012, 2013, and 2014 (1, 2, 3, and 4 Week Predictions)

We also developed regression, boosting and SVM-based prediction models for peak ILI values and the peak week for the flu season. Again, SVM gave the best performance, with the model's results summarized below. Figure 4 shows the predictions for the peak ILI values at national and regional levels using SVM.

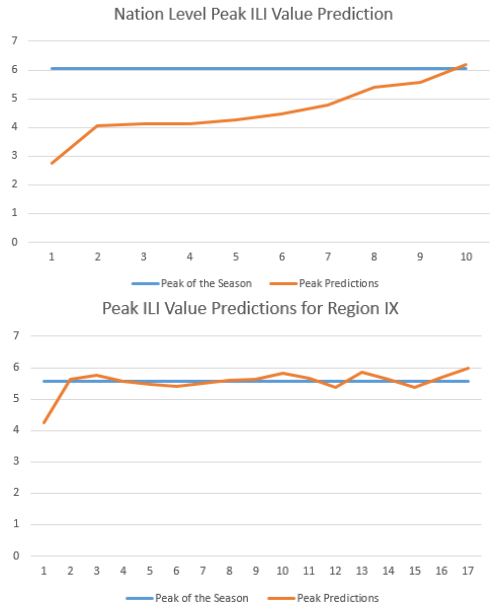


Figure 4. Peak ILI Value Predictions at National and Regional Levels

Figure 5 shows the predictions for the timing of peak ILI at national and regional levels from the SVM model. These predictions give the number of weeks from the current week that the peak ILI is expected to occur.

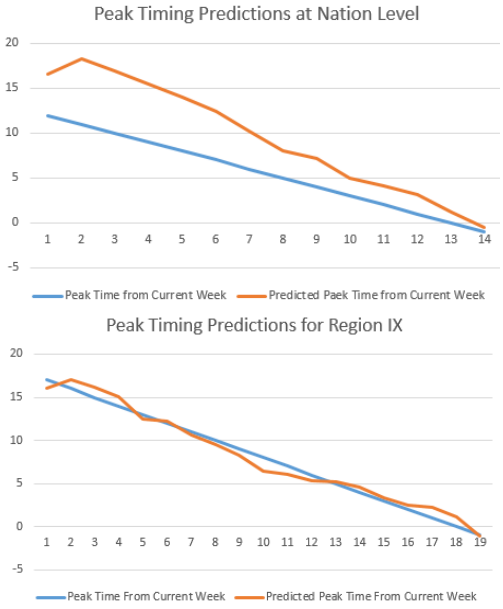


Figure 5. Peak Timing Predictions at National and Regional Levels

The 1-, 2-, 3-, and 4-week look-ahead predictions at the regional level are presented in Figure 6. The results are similar to those at the national level presented in Figure 3.

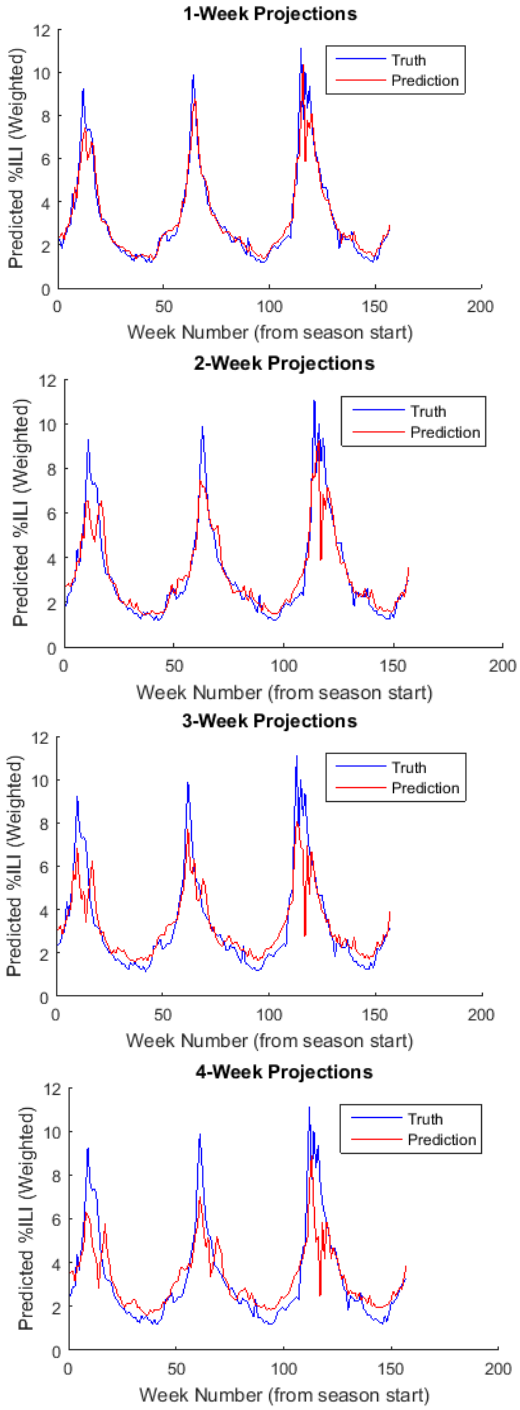


Figure 6. Predicted Percent ILI for HHS Region 6 for 2012 through 2014

We also developed models for predicting the season start time. The season start time is defined as the first time in the flu season when the ILI values are above a baseline threshold for three consecutive weeks. The baseline threshold is the mean ILI plus two standard deviations from all non-influenza weeks in the last 3 years. Non-influenza weeks are periods of two or more consecutive weeks in which each week accounts for less than 2% of the season's total number of influenza-infected specimens[1]. The actual and predicted season start times at region level, as a number of weeks after the current week, are shown in Figure 7, along with the predictions for season peak ILI and week.

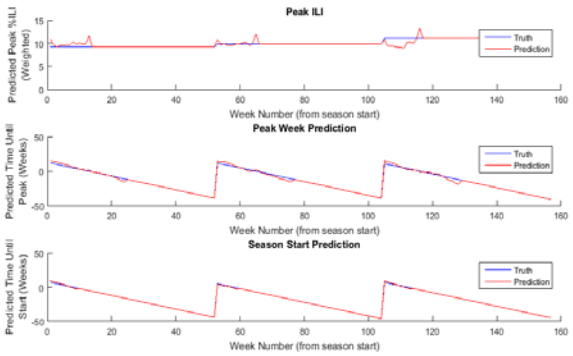


Figure 7. Predicted Peak and Season Start Variables for HHS Region 6 for 2012 through 2014

Based on the residuals (Figure 8) and confidence intervals (Figure 9), the prediction model performance is good.

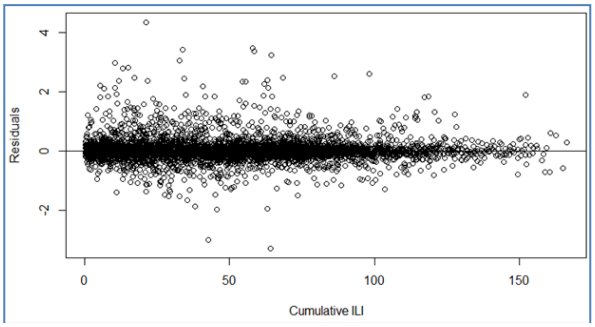


Figure 8. Model Residuals for Cumulative ILI

[1] http://www.cdc.gov/flu/weekly/overview.htm#Viral

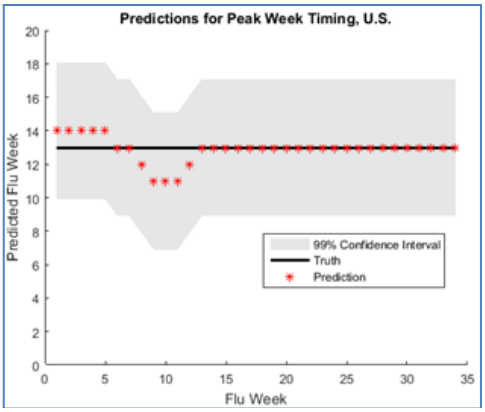


Figure 9. Confidence Interval of the Peak Week vs Actual Peek Week

Insights into ILI pattern gained during the predictive model development are:

1) HHS region and season week were significant, indicating not only the seasonal nature of flu but differences between the 10 HHS regions and the national level flu patterns.
2) ILI value, as well as the square and cube of ILI values were significant factors in the model, indicating the non-linear nature of the model.
3) Percent of ILI value change in the last 2 and 3 weeks were significant factors, indicating that trend is an important influencing factor.
4) Last week's and two week's prior ILI values were significant, indicating the time series nature of the flu.
5) The virus ratio (positive virus A to B ratio) was a significant factor, indicating that flu patterns of virus types A and B are different.
6) Weather had an insignificant effect on the ILI values. Maximum temperature had a very small influence while the remaining weather factors had no significant influence. Weather variables were excluded from the final model. Previous studies have identified that humidity is slightly correlated with ILI intensity [17]. We hypothesize that abstracting environmental data at HHS and national levels is masking out the environmental effects. In the IPAS Phase II effort, we will make local health center level ILI predictions and will re-evaluate the significance of environmental variables on ILI activity.
7) The resulting regression model had R-square of around 90% indicating a very good model fit [18].

4. IPAS Disease Management Application

Once prediction models have been developed, the historical data and applicable predictions is portrayed to stakeholders in a meaningful way. Although the prediction models focused on ILI, the IPAS Application enables the exploration of other diseases in general. The IPAS Application is intended to provide this functionality through three primary views:

1) Situational Awareness View
2) Explore and Analyze View
3) Predictions View

In addition to the data previously described for use by the prediction modeling, the IPAS Application also aggregates data gathered through a Natural Language Processing (NLP) Information Extraction Engine. This component is used to extract relevant meaningful information such as disease outbreaks from unstructured data and render it in a format useful for analysis [14, 15]. The flexibility of this tool enables the incorporation of disease, syndromic, and zoonotic details from news feeds, public health reports, and medical publications.

4.1. Situational Awareness View

The situational awareness view of the IPAS application provides the high level overview of the global disease infections that have occurred in the past (Figure 10). This view is designed to support epidemiologists, and public health analysts to:

1) View the history of disease at a particular region (temporal perspective of a disease in a region).
2) Determine diseases occurring in a region (spatial perspective of a region).
3) Observe spread or migration of disease (temporal perspective of a disease).

The situational awareness view consists of three different visual interfaces which corresponds to the three main dimensions of the epidemic data: (1) disease type, (2) location, and (3) time. All the different components of the situational awareness view are integrated to provide the user with brushing and linking features. The idea of brushing and linking is to facilitate visual analytics and let the end user choose the focus disease, location, or time and let the display change accordingly. Interactive changes made in one visualization are automatically reflected in other visualizations. Connecting multiple visualizations through interactive linking and brushing provides the ability to the end users to creatively drill-down, explore and analyze the data. The disease incidents chart on the top left side provides the histogram of all diseases, with the length of the chart representing the number of reported cases of that disease (Figure 10).

Figure 10. Situational Awareness View of IPAS Application

The spatial distribution map on the top helps to understand the location of the selected disease outbreak. The location of the bubble on the map shows where exactly the cases are recorded and the color corresponds to the disease color on the disease incidents map on the top left. The size of the bubble shows the number of cases recorded; i.e., the bigger the bubble, the higher the number of cases recorded. Along with the bubble, the countries are color coded based on the relative distribution of the selected disease(s) among countries. The user can get additional details about each bubble by hovering the mouse over the bubble. The time line chart on the bottom shows the temporal distribution of the selected disease(s). The X-axis shows the year in which the cases are recorded and the Y-axis represents the number of cases recorded. The screenshot given below shows how filters can be applied to different interfaces of the situational awareness view. It shows the cases of whooping cough all over the world between 2010 and 2011 (Figure 11).

Figure 11. Filtering and Focusing in Situation Awareness View

4.2. Explore and Analyze View

The Explore & Analyze view of the IPAS application helps to drilldown into the data to perform exploratory analysis, causal analysis, and hypothesis testing. This view will assist analysts and epidemiologists to identify factors influencing / affecting / correlating with disease intensity such as:

1) Temporal Correlation Analysis and Visualization (Is the temporal pattern of a disease correlated with weather patterns, school schedule, vector prevalence?).
2) Spatial Correlation Analysis and Visualization (Is the spatial pattern of the disease correlated with spatial coordinates of vectors, contaminated water, contaminated air?). This functionality will be completed in the next phase of IPAS implementation.
3) Hypothesis Testing (Is the pattern of the disease different before and after a milestone event?). This functionality will be completed in the next phase of IPAS implementation.

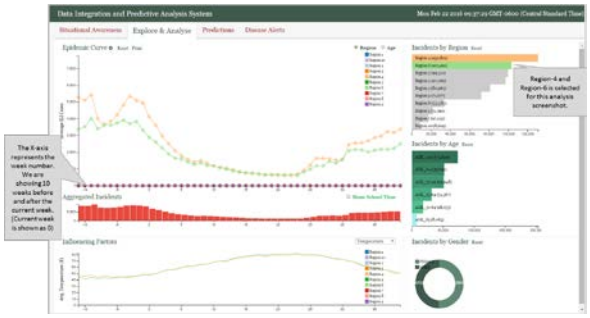


Figure 12. Explore & Analyze View of IPAS

One of the key challenges in building the predictive model is to find the factors that influence the disease pattern. This view helps in gaining the insights required to develop predictive models. The screenshot in Figure 12 shows US flu data by HHS regions.

The "Epidemic curve" shown in Figure 12 gives the graphical display of the number of incidence cases in an outbreak/epidemic, plotted over time. Each line in the chart corresponds to a HHS region which are differentiated using different colors. The bar chart "Aggregated Incidents" on the bottom (red) of the Epidemic curve shows the aggregated number of cases in the selected regions. When no region is selected, it gives the aggregated number of cases for the US. This chart helps to filter the time focus by lasso select and it also shows school open and break times. In the case of most endemic infectious diseases like flu, schools play a major role in transmitting the disease; overlaying the school information on top of the epidemic curve will help an epidemiologist understand that phenomenon. In order to see the school timing, enable the checkbox "Show School Time" on the top right corner of "Aggregated Incidents" (Figure 13).

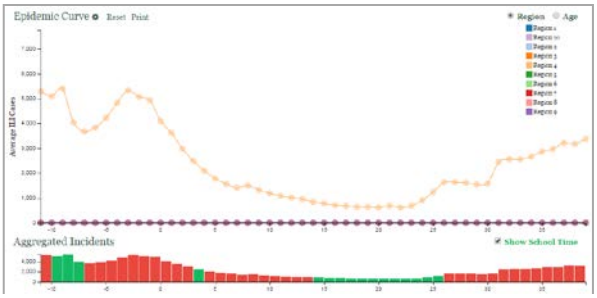


Figure 13. Comparing Epidemic Curve with School Calander

Similar to comparing the epidemic curve between different regions, this interface supports the comparison of the epidemic curve between age groups, gender, and other demographic factors by selecting associated filters and selection (Figure 12). The first row chart of the right side shows the incidents of disease by region (Figure 13). It helps to identify the most affected and least affected regions. It also helps to compare disease patterns in different regions. The second row chart shows the incidents by age groups. This chart helps to find the most affected age group and the least affected age group. The pie chart on the bottom right side shows the incidents by gender (Figure 12).

The "Influencing Factors" chart helps to understand the various factors that influence/affect/correlate with the disease intensity. Figure 14 shows various factors that can be displayed on the chart. For example, comparing the weather patterns and the epidemic curve helps to understand whether a weather pattern has any correlation with the epidemic curve. All the charts in the Explore & Analyze view are integrated to provide the brushing and linking feature.

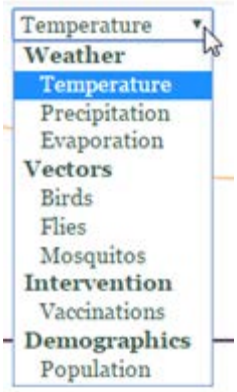


Figure 14. Disease Influencing Factors

4.3. Prediction View

The Prediction View of the IPAS application (Figure 15) helps to explore the results of the analytical and machine learning models about the spread of diseases (Section 3). This view shows the prediction results in an intuitive fashion. It supports epidemiologists and healthcare officials to analyze:

1) Expected disease pattern in the next few weeks and months.
2) Peak timing and intensity.

This view also assists the users in taking prophylaxis decisions based on advanced knowledge. The filters shown on the left side enables the selection of disease(s) and regions of interest (Figure 15).

Figure 15. Prediction View of IPAS

Based on the selection, the geospatial map on the top right highlights the regions and the prediction timeline chart on the bottom shows the actual progression of the selected diseases in the past 26 weeks and the prediction for next 26 weeks. The actual progression is shown to facilitate the comparison of the accuracy of the prediction model. The predicted peak week is also highlighted in the chart. Information about the current values and predicted peak values are also shown in the information box.

5. Summary

Advance knowledge about the location, timing, and intensity of infectious diseases will help public health stakeholders in taking proactive disease containment and management efforts. IPAS is motivated to provide predictive modeling infrastructure to support this important public health functionality. IPAS prediction modelling was focused on Influenza Like Illness (ILI) in initial effort and the ILI prediction results were presented in this paper.

IPAS supports comprehensive, end-to-end support for exploratory analysis, temporal correlation analysis,

and prediction. The features of IPAS system that support the different stages in epidemiological data collection, integration, and analysis were presented in this paper.

In future work we plan to enhance prediction by including Google Flu Trend data [21]. This data is provided by Google, and represents the number of searches related to ILI. As more social media data becomes available from HealthTweets, we intend to continue to develop our prediction model to include these values appropriately. We intend to implement other prediction methodologies, such as non-linear regression, in addition to the SVM model currently employed. Additionally, we intend to support local healthcare at the service provider level, and state and international level ILI predictions.

6. Acknowledgements

This work was supported by the Defense Health Program (DHP) under Small Business Innovative Research (SBIR) Contract No. W81XWH-15-C-0158. We would like to acknowledge the guidance and inputs provided by Mr. Robert Huffman and Mr. Jeffrey Morgan. In the next Phase of the IPAS effort, the system will be piloted at Defense Threat Reduction Agency's (DTRA) Biosurveillance Ecosystem (BSVE) [22].

7. References

[1] Stephen S Morse, Jonna A K Mazet, Mark Woolhouse, Colin R Parrish, Dennis Carroll, William B Karesh, Carlos Zambrana-Torrelio, W Ian Lipkin, and Peter Daszak, "Prediction and prevention of the next pandemic zoonosis," Lancet. 2012 December 1; 380(9857): 1956–1965. doi:10.1016/S0140-6736(12)61684-5.

[2] Sarah H. Olson, Corey M. Benedum, Sumiko R. Mekaru, Nicholas D. Preston, Jonna A.K. Mazet, Damien O. Joly, John S. Brownstein, "Drivers of Emerging Infectious Disease Events as a Framework for Digital Detection," Emerging Infectious Diseases, Vol. 21, No. 8, August 2015.

[3] Corley CD, Pullum LL, Hartley DM, Benedum C, Noonan C, Rabinowitz PM, et al. (2014) Disease Prediction Models and Operational Readiness. PLoS ONE 9(3): e91989. doi:10.1371/journal.pone.0091989.

[4] H. Hethcote, "The Mathematics of Infectious Diseases," *SIAM Review*, vol. 42, no. 4, p. 599–653, 2000.

[5] R. J. D. Tebbens, "A Dynamic Model of Poliomyelitis Outbreaks: Learning from the Past to Help Inform the Future," *American Journal of Epidemiology*, vol. 162, no. 4, pp. 358-372, 2005.

[6] B. T. Mayer, J. N. S. Eisenberg, C. J. Henry, G. M. Gomes, E. L. Ionides and J. S. Koopman, "Successes and Shortcomings of Polio Eradication: A Transmission Modeling Analysis," *American Journal of Epidemiology*, vol. 177, no. 11, pp. 1236-45, 2013.

[7] "Ebola modeling workshop at Georgia Tech", http://dx.doi.org/10.6084/m9.figshare.1301267

[8] Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R (2015) Flexible Modeling of Epidemics with an Empirical Bayes Framework. PLoS Comput Biol vol. 11, no. 8: e1004382. doi:10.1371/ journal.pcbi.1004382

[9] Wolfe, N., Daszak, P., Kilpatrick, A. M., and Burke, D. "Bushmeat hunting, deforestation, and prediction of zoonotic disease emergence." *Emerging Infectious Diseases*, Vol. 11 No. 12, 2005

[10] Jones, K., Patel, N., Levy, M., Storeygard, A., Balk, D., Gittleman, J., and Daszak, P. "Global trends in emerging infectious diseases." *Nature*, Vol. 451, 2008.

[11] Paull, S., Song, S., McClure, K., Sackett, L., Kilpatrick, A. M., and Johnson, P. "From superspreaders to disease hotspots: linking transmission across hosts and space." *Frontiers in Ecology and the Environment*, Vol. 10, no. 2. March 2012.

[12] HealthMap, "HealthMap," [Online]. Available: http://www.healthmap.org/en/. [Accessed 08 March 2016].

[13] VectorMap, "Know the vector, Know the threat," Walter Reed Biosystematics Unit, [Online]. Available: www.vectormap.org/. [Accessed 08 March 2016].

[14] M. Erraguntla, S. Ramachandran, C.-N. Wu and R. J. Mayer, "Avian Influenza Data mining Using Environment, Epidemiology, and Etiology Surveillance and Analysis Toolkit (E3SAT)," in Hawaii International Conference on System Sciences, Hawaii.

[15] S. Ramachandran, M. Erraguntla, R. Mayer and P. Benjamin, "Data Mining in Military Health Systems – Clinical and Administrative Applications," in IEEE Conference on Automation Science and Engineering, 2007.

[16] CDC, "FluView," CDC, [Online]. Available: http://www.cdc.gov/flu/weekly/fluviewinteractive.htm. [Accessed 08 March 2016].

[17] Wan Yang, Marc Lipsitchb, and Jeffrey Shamana, "Inference of seasonal and pandemic influenza transmission dynamics," PNAS, March 2015, vol. 112, no. 9, pp. 2723–2728.

[18] E. J. Pedhazur, 1997, Multiple Regression in Behavioral Research, 3rd ed, Harcourt, Inc: Troy, MO.

[19] M. Erraguntla, L. May, B. Gopal and R. J. Mayer, "Open Data Sources Based Biovigilance," in International Conference on Artificial Intelligence, Las Vegas, 2012.

[20] P. Benjamin, K. Madanagopal, M. Erraguntla, and D. Corlette. 2016. “Distributed Information Gathering, Exploration and Sense-making Toolkit (DIGEST),” Accepted for publication in ICAI'16 - The 2016 International Conference on Artificial Intelligence, Las Vegas.

[21] Google, "The Next Chapter for Flu Trends," [Online]. Available: http://googleresearch.blogspot.com/2015/08/the-next-chapter-for-flu-trends.html. [Accessed 08 March 2016].

[22] DTRA, "Biosurveillance Ecosystem (BSVE)," DTRA, 08 March 2016. [Online]. Available: http://www.dtra.mil/Portals/61/Documents/bsve-fact-sheet_draft_05-01-2014_pa-cleared-distro-statement.pdf. [Accessed 08 March 2016].

[23] Delen, D. 2015. "Real-World Data Mining: Applied Business Analytics and Decision Making". Upper Saddle River, New Jersey: Financial Times Press (Pearson Company).